

Fairness-Aware Optimization in Logistic Regression: Empirical Trade-Offs Across Healthcare and Socioeconomic Domains

David Oluwabukunmi Shodipo and Blessing Ogbuokiri

Responsible and Applied Machine Laboratory (RAML Lab),
Department of Computer Science,
Brock University, Canada

Abstract

With the increased use of machine learning predictive models, there is a growing need to ensure algorithmic fairness and balance out inequality in predictive errors across various application domains, such as health care and income prediction. To make machine learning algorithms more useful and applicable in everyday settings, the issue of algorithmic fairness and the differences in error rates across groups has gained importance. While numerous efforts have been made to detect bias, very few studies consider the effect of incorporating explicit constraints into an optimization framework on the training process. This paper examines how the fairness regularization parameter (λ) affects the predictive and fairness characteristics of logistic regression models trained via gradient descent. The experimental results were obtained using two datasets: the UCI Adult Income dataset and the UCI Heart Disease dataset. The experiment uses either demographic parity or an equal opportunity regularizer, along with cross-entropy loss for each model. To evaluate the results, the demographic parity difference and the equal opportunity difference are considered in addition to accuracy and AUC. The results show that fairness interventions are highly domain dependent: the same regularization strategy substantially improves fairness in large socioeconomic datasets but provides limited benefits in small healthcare datasets. In the Adult Income dataset, higher tuning parameter values led to lower fairness gaps and slightly increased accuracy; in the Heart Disease dataset, the gap improvement is smaller but accompanied by decreased performance. The findings highlight that fairness-aware optimization should be treated as a context-sensitive design choice rather than a universally transferable solution.

Introduction

Machine learning models are now central to decision-making in domains that directly affect human lives, including healthcare diagnostics, credit scoring, hiring, and criminal justice (Mehrabi et al. 2022; Sufian et al. 2024; Kpatcha 2025). These systems learn patterns from historical data to make predictions about future outcomes. However, historical data frequently encode societal biases, such as income disparities along gender lines or unequal access to healthcare across demographic groups. When models are trained

solely to maximize predictive accuracy, they risk learning and amplifying these existing inequalities, producing systematic disparities in results between demographic groups (Mehrabi et al. 2022).

This problem, often called *algorithmic bias*, has attracted significant research attention. Fairness in machine learning broadly refers to the goal of ensuring that predictive models do not produce unjustifiably different outcomes for individuals based on sensitive attributes such as sex, race, or age. Multiple formal definitions of fairness have been proposed, including demographic parity (Dwork et al. 2012; Hardt, Price, and Srebro 2016). Each definition captures a different notion of equitable treatment and may be more or less appropriate depending on the application context.

In healthcare settings, fairness is particularly consequential. A model that systematically misses diagnoses in one group causes real-world harm, not merely a statistical disparity. Algorithmic bias in cardiovascular diagnostics has been documented to produce differential performance between sex and racial groups (Sufian et al. 2024). Outside of healthcare, income prediction systems that assign positive outcomes more often to a demographic group reflect and reinforce social inequalities in economic opportunity.

Importantly, fairness interventions are not purely technical mechanisms. Different fairness definitions correspond to different ethical priorities, and the choice of optimization objective can redistribute predictive errors across demographic groups in ways that affect access to healthcare, employment, or economic opportunity. Understanding how fairness regularization behaves across domains is therefore important not only for model performance, but also for responsible AI deployment.

Most machine learning pipelines treat fairness as an afterthought, evaluating bias only after a model has been trained for accuracy. This post-hoc approach limits the ability to understand how optimization choices, including learning rate, training duration, and regularization strength, interact with fairness outcomes. Although fairness-aware learning has been widely studied, less attention has been given to how fairness regularization behaves across domains with different dataset sizes, feature structures, and ethical risk profiles.

This study makes the following contributions:

- It implements regularized fairness logistic regression with gradient descent, incorporating both demographic parity (DP) and equal opportunity (EO) as differentiable regularization terms.
- It conducts a systematic empirical analysis of fairness–accuracy trade-offs across nine regularization strengths and three epoch budgets.
- It compares a non-healthcare domain (UCI Adult Income) with a healthcare domain (UCI Heart Disease).
- It provides preliminary empirical evidence that fairness interventions may not transfer uniformly across the socioeconomic and healthcare domains studied here, motivating further domain-sensitive fairness evaluation.
- It analyzes optimization dynamics, including loss curves and fairness metric trajectories across epochs.
- It presents practical λ recommendations based on the observed fairness–performance of the Pareto frontiers.
- It investigates random-seed sensitivity using a one-sample t-test and Cohen’s d, showing that fairness metrics require multi-seed evaluation to draw reliable conclusions.

Related Work

Hardt, Price, and Srebro (2016) introduced equalized odds and equal opportunity in supervised learning, proposing a post-hoc correction method to adjust a learned predictor so that true positive rates are equalized across groups. Complementarily, Dwork et al. (2012) formalized demographic parity as requiring equal positive prediction rates across sensitive groups. The tension between these definitions, including the difficulty of simultaneously satisfying demographic parity, equal opportunity, and predictive parity, except in degenerate cases, has since been extended by Defrance and De Bie (2025), who identified maximal combinations of fairness definitions that can be satisfied simultaneously.

A growing body of work embeds fairness directly into the training objective rather than applying corrections post-hoc. Rezaei et al. (2020) proposed a robust log-loss classifier derived from distributional robustness principles that incorporates fairness criteria into worst-case logarithmic loss minimization. Trotter and Chen (2024) explored fairness–accuracy trade-offs in binary classification using modified binary cross-entropy loss functions, an approach closely related to the present study, and observed that stronger fairness constraints produce steeper accuracy declines in datasets with large historical biases.

Gradient-based fairness methods are especially relevant here. Maheshwari and Perrot (2023) introduced FairGrad, a gradient descent method that iteratively reweights group-specific gradients based on whether groups are advantaged. Unlike the scalar regularization parameter λ used in this study, FairGrad adapts weights dynamically during training. Buyl, Defrance, and De Bie (2024) introduced a unified framework of differentiable fairness regularization terms that can be integrated with automatic differentiation libraries, while Taskesen et al. (2020) proposed a distributionally robust logistic regression model with an unfairness penalty. Rychener, Taskesen, and Kuhn (2024) further

proved that fairness-regularized prediction loss admits unbiased gradient estimators when unfairness is measured by squared L^2 distance or squared maximum mean discrepancy.

Recent work also complicates the assumption that fairness and accuracy must always be in trade-off. Blum and Stangl (2019) showed that equal opportunity constraints combined with empirical risk minimization can recover the Bayes-optimal classifier when training data is biased. Hort et al. (2024) presented search-based repair methods that increase both fairness and accuracy in decision-making software. Liang et al. (2024) characterized a fairness–accuracy frontier and showed that a dataset property called group-balance qualitatively determines its shape. This concept directly motivates the λ recommendations reported in this study as empirical operating points on each dataset’s frontier.

In healthcare contexts, the choice of a fairness metric has distinct ethical consequences. Sufian et al. (2024) investigated algorithmic bias in cardiovascular risk prediction and emphasized that false negatives in healthcare represent missed diagnoses, not merely unfavourable statistical outcomes. This motivates the use of an equal-opportunity approach for the Heart Disease dataset. Sánchez-Marqués, Garcia, and Sanchez (2026) examined fairness constraints in glioma grade prediction and found that different mitigation methods can have divergent effects, suggesting that in-processing regularization requires empirical domain-specific evaluation.

Finally, fairness evaluation is sensitive to data splits and subgroup composition (Rahman and Ogbuokiri 2026; Martin and Ogbuokiri 2026). Brzezinski et al. (2025) showed that measures such as equal opportunity are more sensitive to changes in class imbalance than accuracy equality. This motivates the random seed analysis conducted here, which tests whether a single reference split is representative of the multi-seed distribution.

Unlike prior work that focuses primarily on proposing new fairness algorithms, this study emphasizes empirical analysis of optimization behaviour under fairness regularization across domains with substantially different characteristics. Specifically, it investigates how regularization strength influences fairness stability, predictive performance, and random-seed sensitivity in both healthcare and socioeconomic settings.

Methodology

This study uses logistic regression because its linear structure allows clearer interpretation of the effects of fairness regularization without the additional complexity introduced by deep neural architectures. Demographic parity and equal opportunity are selected because they represent two widely used but ethically distinct notions of fairness: outcome parity and diagnostic sensitivity parity. The goal of this study is not to propose a new fairness algorithm, but to systematically analyze optimization behavior under fairness-aware regularization.

Datasets and Pre-processing

The Adult Income dataset contains 48,790 observations and 15 features describing socio-economic attributes, including age, education level, occupation, marital status, hours worked per week, capital gain, capital loss, and native country. The prediction task is binary: determine whether an individual's annual income exceeds \$50,000. The sensitive attribute is sex, with the male group treated as privileged and the female group as unprivileged.

In the Adult Income dataset, the unprivileged (female) group comprises 16,176 individuals with a positive-label rate of 10.9%, while the privileged (male) group comprises 32,614 individuals with a positive-label rate of 30.4%. In the Heart Disease dataset, the unprivileged (female) group comprises 97 individuals with a positive-label rate of 22.7%, while the privileged (male) group comprises 206 individuals with a positive-label rate of 37.4%.

The Heart Disease dataset contains 303 observations and 14 clinical features, including age, resting blood pressure, serum cholesterol, fasting blood sugar, resting ECG results, maximum heart rate, and exercise-induced angina. The prediction task is binary: determine the presence of heart disease. The sensitive attribute is sex, encoded as 0 for female and 1 for male, with the male group treated as privileged. This dataset is substantially smaller than the Adult dataset, making it more susceptible to random seed variation in fairness metrics.

For both datasets, string columns were stripped of whitespace, string-encoded missing values were replaced with proper missing entries, duplicate rows were removed, numeric features were imputed using medians and standardized with z -score scaling, categorical features were imputed using the most frequent value and one-hot encoded, and the sensitive attribute was excluded from the feature matrix and encoded separately. The datasets were split into 70% training and 30% test sets, with random seed 2026 used for baseline experiments.

Fairness Metrics

Let $\hat{Y}$ denote the predicted label, $A \in \{0, 1\}$ the sensitive attribute (0 = unprivileged, 1 = privileged), and Y the true label. Demographic parity difference is defined as:

$$DP = P(\hat{Y} = 1 | A = 0) - P(\hat{Y} = 1 | A = 1). \quad (1)$$

Demographic parity measures whether the positive prediction rate is equal across groups. A negative value indicates that the privileged group receives positive predictions more frequently.

Equal opportunity difference is defined as:

$$EO = \text{TPR}(A = 0) - \text{TPR}(A = 1). \quad (2)$$

Equal opportunity measures whether the true positive rate is equal across groups. This metric is especially important in healthcare, where a low true positive rate for one group corresponds to more missed diagnoses.

Fairness-Regularized Optimization

The baseline model is a logistic regression classifier trained using scikit-learn's `LogisticRegression` with standard pre-processing and no fairness regularization. The

fairness-regularized model uses a logistic regression predictor:

$$\hat{p}_i = \sigma(w^\top x_i + b), \quad \sigma(z) = \frac{1}{1 + e^{-z}}. \quad (3)$$

The standard prediction loss is binary cross-entropy:

$$\mathcal{L}_{CE}(w, b) = -\frac{1}{n} \sum_{i=1}^n [y_i \log \hat{p}_i + (1 - y_i) \log(1 - \hat{p}_i)]. \quad (4)$$

Fairness is incorporated by adding a differentiable penalty:

$$\mathcal{L}_{\text{total}}(w, b) = \mathcal{L}_{CE}(w, b) + \lambda R_{\text{fair}}(w, b), \quad (5)$$

where $\lambda \geq 0$ controls the strength of the fairness penalty. When $\lambda = 0$, training reduces to accuracy-only logistic regression. As λ increases, the model is increasingly pressured to reduce group disparities.

For demographic parity regularization, the penalty is the squared difference in average predicted probabilities between groups:

$$R_{DP} = \left(\frac{1}{n_1} \sum_{i:a_i=1} \hat{p}_i - \frac{1}{n_0} \sum_{i:a_i=0} \hat{p}_i \right)^2. \quad (6)$$

For equal opportunity regularization, the penalty is restricted to individuals with true positive outcomes:

$$R_{EO} = \left(\frac{1}{n_1^+} \sum_{i:a_i=1, y_i=1} \hat{p}_i - \frac{1}{n_0^+} \sum_{i:a_i=0, y_i=1} \hat{p}_i \right)^2. \quad (7)$$

The model is trained via gradient descent using $\theta_{t+1} = \theta_t - \eta \nabla_{\theta} \mathcal{L}_{\text{total}}$, where $\theta = (w, b)$ and $\eta = 0.1$.

Algorithm

The full fairness-regularized training procedure is formalized in Algorithm 1.

Algorithm 1 Fairness-Regularized Gradient Descent Training

Require: Dataset $\mathcal{D} = \{(x_i, y_i, a_i)\}_{i=1}^n$, $\lambda \geq 0$, learning rate η , max epochs T

Ensure: Trained weights w , bias b

```
1: Preprocess  $\mathcal{D}$ : impute, scale numerics, one-hot encode
   categoricals
2: Encode sensitive attribute  $a \in \{0, 1\}$ 
3: Split  $\mathcal{D}$  into  $\mathcal{D}_{\text{train}}$  and  $\mathcal{D}_{\text{test}}$  (70/30)
4: Initialise  $w \leftarrow \mathbf{0}$ ,  $b \leftarrow 0$ 
5: for epoch  $t = 1, 2, \dots, T$  do
6:   Compute predicted probabilities:  $\hat{p}_i = \sigma(w^\top x_i + b)$ ,
   where  $\sigma(z) = 1/(1 + e^{-z})$ 
7:   Compute prediction loss (cross-entropy)  $\mathcal{L}_{\text{CE}}$ 
8:   if DP then
9:      $R_{\text{fair}} = (\mathbb{E}[\hat{p} \mid a = 1] - \mathbb{E}[\hat{p} \mid a = 0])^2$ 
10:  else
11:     $R_{\text{fair}} = (\mathbb{E}[\hat{p} \mid a = 1, y = 1] - \mathbb{E}[\hat{p} \mid a = 0, y = 1])^2$ 
12:  end if
13:  Compute total loss:  $\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{CE}} + \lambda \cdot R_{\text{fair}}$ 
14:  Compute gradient  $\nabla_{\theta} \mathcal{L}_{\text{total}}$  w.r.t.  $(w, b)$ 
15:  Update:  $w \leftarrow w - \eta \cdot \nabla_w \mathcal{L}_{\text{total}}$ ;  $b \leftarrow b - \eta \cdot \nabla_b \mathcal{L}_{\text{total}}$ 
16:  Record accuracy, AUC, DP gap, EO gap
17: end for
18: return  $w$ ,  $b$ , and recorded metrics
```

Experimental Setup

Experiments evaluate $\lambda \in \{0.0, 0.1, 0.3, 0.5, 0.6, 0.7, 0.8, 0.9, 1.0\}$, epoch budgets of 20, 30, and 50, and two fairness objectives: DP and EO. Accuracy, AUC, DP gap, and EO gap are recorded per epoch and at convergence. A multi-seed evaluation uses 30 seeds to assess sensitivity, while a reference seed of 2026 is used for baseline comparisons. All experiments were implemented in Python using NumPy for gradient computations and scikit-learn for pre-processing and evaluation metrics.

Results

Across random seeds, fairness metrics showed substantially greater variability than accuracy metrics, underscoring the importance of reporting multi-seed uncertainty in addition to point estimates.

Baseline Disparities

Table 1 presents baseline performance and fairness metrics for both datasets with no fairness regularization. Negative fairness values indicate that the privileged group receives higher positive prediction rates or higher true positive rates.

For the Adult Income dataset, the baseline model achieves accuracy of 0.8540. The demographic parity difference of -0.1769 indicates that male individuals receive positive income predictions at a rate 17.7 percentage points higher than female individuals. The equal-opportunity difference of -0.0999 indicates that the model is approximately 10 percentage points more effective at identifying true high-income earners among males than among females.

Dataset	Accuracy	DP Diff.	EO Diff.
Adult Income	0.8540	-0.1769	-0.0999
Heart Disease	0.7802	-0.1279	-0.2236

Dataset	Precision	Recall	F1
Adult Income	0.7414	0.5990	0.6627
Heart Disease	0.6923	0.6000	0.6429

Table 1: Baseline accuracy, fairness, and classification metrics ($\lambda = 0$).

For the Heart Disease dataset, the model achieves accuracy of 0.7802. Although the demographic parity difference (-0.1279) is smaller than for the Adult dataset, the equal opportunity difference (-0.2236) is substantially larger. In a healthcare context, this means that patients with heart disease in the unprivileged group are substantially more likely to receive a missed diagnosis.

Precision, recall, and F1-score (Table 1) provide additional context beyond accuracy. For both datasets, recall is lower than precision, indicating that positive cases (high income for Adult Income, presence of heart disease for Heart Disease) are missed more often than they are falsely flagged. This is particularly consequential for the Heart Disease dataset: a recall of 0.6000 means the baseline model fails to identify 40% of patients who actually have heart disease, independent of demographic group. The equal opportunity difference of -0.2236 reported above therefore compounds an already substantial baseline miss rate rather than describing a disparity in an otherwise well-performing model.

Effect of Fairness Regularization

Figure 1 summarizes the λ sweep across both datasets and both fairness criteria. In the figure, fairness is reported as an *absolute* gap (i.e., $|\text{DP}|$ and $|\text{EO}|$) on the test set for ease of comparison across settings.

Recommendation rule. For each dataset and fairness objective, we select λ^* by minimizing the corresponding absolute fairness gap on the test set subject to an accuracy budget relative to the unconstrained model:

$$\lambda^* = \arg \min_{\lambda \in \Lambda} |\Delta_{\text{fair}}(\lambda)| \quad \text{s.t.} \quad \text{Acc}(\lambda) \geq \text{Acc}(0) - 0.01, \quad (8)$$

where Δ_{fair} is either DP or EO and the budget allows at most a 1 percentage-point absolute accuracy drop. If multiple λ values satisfy the constraint and attain similar fairness gaps, we choose the smaller λ (a simpler intervention). If no $\lambda > 0$ yields a meaningful reduction in $|\Delta_{\text{fair}}|$ under the budget, we recommend $\lambda = 0$.

For the Adult dataset under DP regularization, $|\text{DP}|$ decreases as λ increases, reaching its lowest values at higher λ (with a best operating point near $\lambda \approx 0.8$). Under EO regularization on the Adult dataset, $|\text{EO}|$ similarly decreases with λ , with the strongest reduction occurring near $\lambda \approx 1.0$.

Note that Table 1 reports a scikit-learn baseline model, whereas Figure 1 reports results from our gradient-descent

Setting	Recommended λ	Accuracy Impact
Adult + DP	≈ 0.8	Slight increase
Adult + EO	≈ 1.0	Stable
Heart + DP	≈ 0.5	Slight decrease
Heart + EO	0.0	Stable

Table 2: Recommended λ values and observed effects.

training implementation (including the $\lambda = 0$ point). As a result, absolute metric values at $\lambda = 0$ need not match exactly across the table and figure; the comparison of interest is the within-implementation trend as λ varies.

The Heart dataset behaves differently. Under DP regularization, the strongest improvement occurs around $\lambda = 0.5$, where accuracy decreases only slightly relative to the $\lambda = 0$ baseline (Table 1). Beyond $\lambda = 0.5$, further increases produce diminishing fairness gains with continued accuracy degradation. Under EO regularization, accuracy, AUC, and fairness metrics remain nearly constant across all tested λ values, so the baseline is preferred because increasing λ provides no meaningful fairness improvement.

Training Dynamics and Cross-Domain Behavior

Training curves were recorded across all λ values and epoch budgets. Across all settings, training is smooth and stable, with no divergence or oscillation. Total loss decreases consistently across epochs, and fairness metrics evolve gradually as λ increases. Increasing the epoch budget from 20 to 30 to 50 improves final metrics, but marginal gains decrease at higher epoch counts, suggesting that 50 epochs is approximately sufficient for convergence at the tested learning rate.

The Adult dataset shows clear λ -dependent separation in both loss and fairness trajectories, especially under DP regularization. The Heart dataset shows weaker separation, particularly under EO regularization, reflecting the limited effect of regularization in a small and constrained dataset. Figure 2 summarizes this cross-domain pattern: Adult Income benefits substantially from regularization, while Heart Disease shows only modest DP improvement.

Random Seed Sensitivity

A one-sample t -test was conducted to determine whether results from the reference seed are statistically representative of the distribution across random seeds. Let $D_i = X_i - X_r$ be the difference between metric values from seed i and the reference seed. The null hypothesis is $H_0 : \mu_D = 0$, with:

$$t = \frac{\bar{D}}{s_D/\sqrt{n}}, \quad \text{Cohen's } d = \frac{\bar{D}}{s_D}. \quad (9)$$

For the Adult dataset, accuracy and AUC show statistically significant differences from the reference seed ($p < 0.05$), but the practical impact is very small (std ≈ 0.0023). Fairness metrics, especially equal opportunity difference, show larger spread and frequent statistical significance, with moderate to large Cohen’s d values.

For the Heart dataset, variability is much larger because of its small size: accuracy standard deviation is approximately 0.039 and fairness standard deviation ranges from

approximately 0.09 to 0.14. Several fairness metrics differ significantly from the reference seed, confirming that a single random split can produce materially misleading fairness conclusions in small datasets. These findings establish that multi-seed evaluation with mean and standard deviation reporting is necessary for reproducible fairness analysis.

Because the Heart Disease dataset is relatively small, fairness estimates should be interpreted cautiously despite multi-seed evaluation.

Discussion

The results demonstrate that fairness regularization introduces a clear and controllable trade-off between predictive performance and group fairness. Rather than treating fairness as a post-hoc evaluation metric, incorporating it directly into the optimization objective allows the model to adjust its behavior during training in a structured and interpretable way.

A key observation is that fairness regularization behaves analogously to classical regularization. When $\lambda = 0$, the model focuses entirely on minimizing prediction error and may exploit biased patterns in the data that correlate sensitive group membership with the target. As λ increases, the fairness term constrains the model, reducing group disparities while preserving most predictive signal. If λ becomes excessively large, the model may underfit by sacrificing legitimate predictive information to satisfy fairness constraints.

The simultaneous improvement of fairness and accuracy at moderate λ values in the Adult dataset is consistent with Blum and Stangl (2019) and Hort et al. (2024), who show that fairness constraints can correct bias rather than simply impose performance cost. For large, feature-rich datasets with substantial historical bias, fairness regularization may redirect the model away from biased shortcuts and toward more generalizable predictive structure.

The most striking finding is the difference in regularization effectiveness between the Adult and Heart datasets. The Adult dataset responds strongly to fairness regularization, with clear and monotonic improvement across a wide range of λ values. The Heart dataset shows limited responsiveness, especially to EO regularization, and exhibits tighter accuracy trade-offs. Dataset size appears central: Adult contains 48,790 observations, while Heart contains only 303. Feature richness also matters, because the Adult dataset offers more alternative predictive signals that can support fairness-compatible classification.

For practitioners, the results suggest three guidelines. First, the choice of fairness metric matters: demographic parity regularization was more effective in these experiments, while equal opportunity remains ethically salient in healthcare because it addresses missed diagnoses. Second, λ should be selected empirically from the observed fairness–accuracy curve rather than chosen by assumption. Third, fairness results should be reported across multiple random seeds, since single-split fairness conclusions can be misleading, especially for small datasets.

Although this study uses logistic regression for interpretability, the fairness regularization framework itself is not

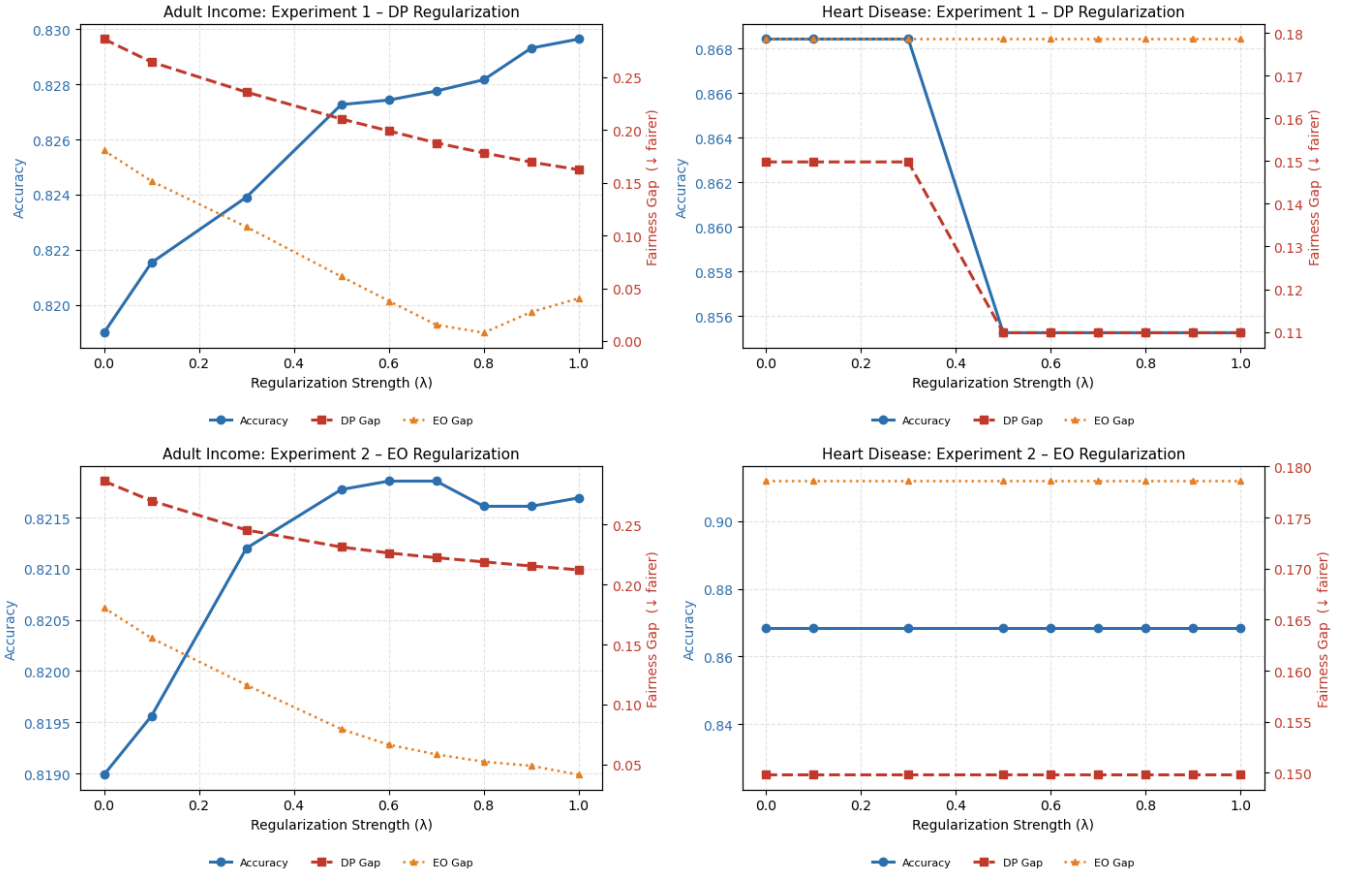


Figure 1: Accuracy–fairness trade-off across λ values for all four dataset–criterion combinations. Accuracy and AUC are shown with the left axis, while absolute fairness gaps are shown with the right axis. Dashed lines mark recommended λ values.

tied to a linear model. The penalties R_{DP} and R_{EO} in Equations (6)–(7) depend only on the predicted probabilities $\hat{p}_i$, not on the functional form that produces them. The same penalties could therefore be added, with no change in structure, to the training objective of other interpretable models that output calibrated probabilities, such as generalized additive models or gradient-boosted trees trained with a differentiable surrogate loss. Extending the approach to black-box models such as deep neural networks would similarly require no modification to R_{DP} or R_{EO} , since $\hat{p}_i = \sigma(w^\top x_i + b)$ could be replaced by the output of any differentiable architecture. Black-box extensions would, however, introduce challenges absent from this study: larger models are more prone to exploiting the fairness penalty during training rather than learning fair representations, gradient interactions between the prediction loss and fairness term become harder to interpret in non-convex, high-dimensional loss landscapes, and the smooth, monotonic fairness–accuracy trade-off observed here for logistic regression may not hold. We view this generalization as a promising direction for future work rather than a structural limitation of the proposed regularization framework.

These findings suggest that fairness interventions cannot be evaluated independently of deployment context. The

same optimization strategy may redistribute predictive errors differently across domains, potentially affecting health-care access, diagnostic sensitivity, or economic opportunity. Improvement in a single fairness metric does not necessarily imply overall fairness, since optimizing one criterion may worsen other forms of disparity. Future fairness-aware systems should incorporate stakeholder-specific definitions of harm rather than relying solely on statistical parity objectives.

Societal Implications and Ethical Considerations

The results of this study have important societal implications because fairness regularization directly affects the distribution of predictive errors across demographic groups. In high-stakes domains, the choice of a fairness objective is not only a technical modelling decision but also an ethical decision about which type of harm to minimize. Demographic parity focuses on reducing differences in positive prediction rates, while equal opportunity focuses on reducing differences in true positive rates. These objectives can lead to different outcomes, and neither should be treated as universally appropriate for every application.

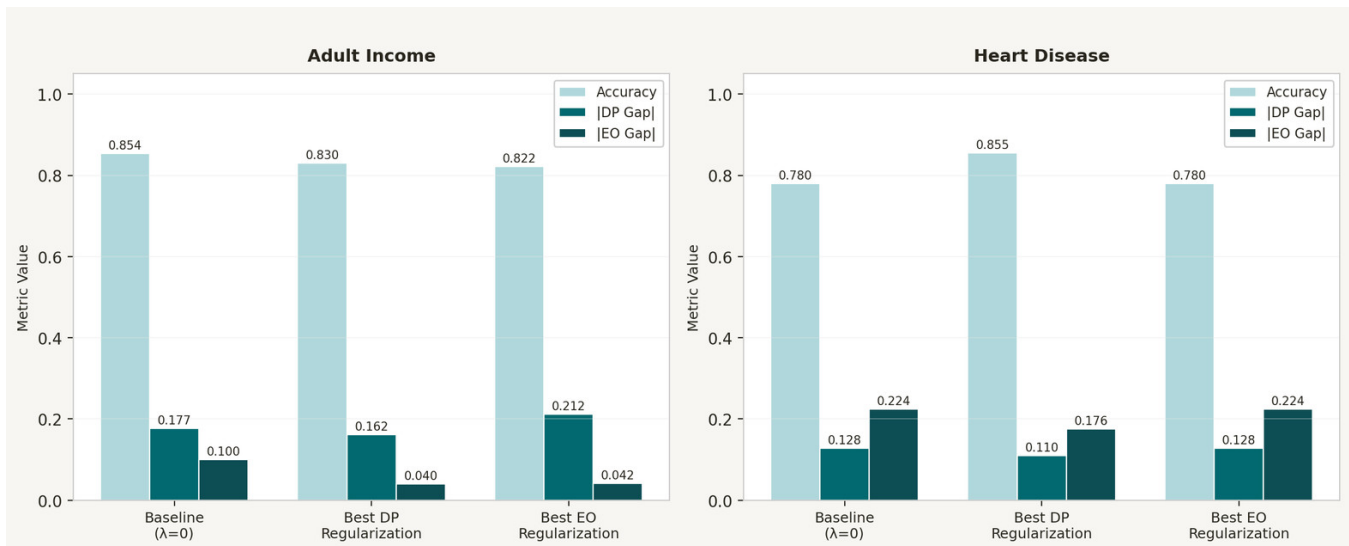


Figure 2: Cross-domain comparison of baseline versus best- λ performance. Adult Income benefits substantially from fairness regularization, while Heart Disease shows more limited gains.

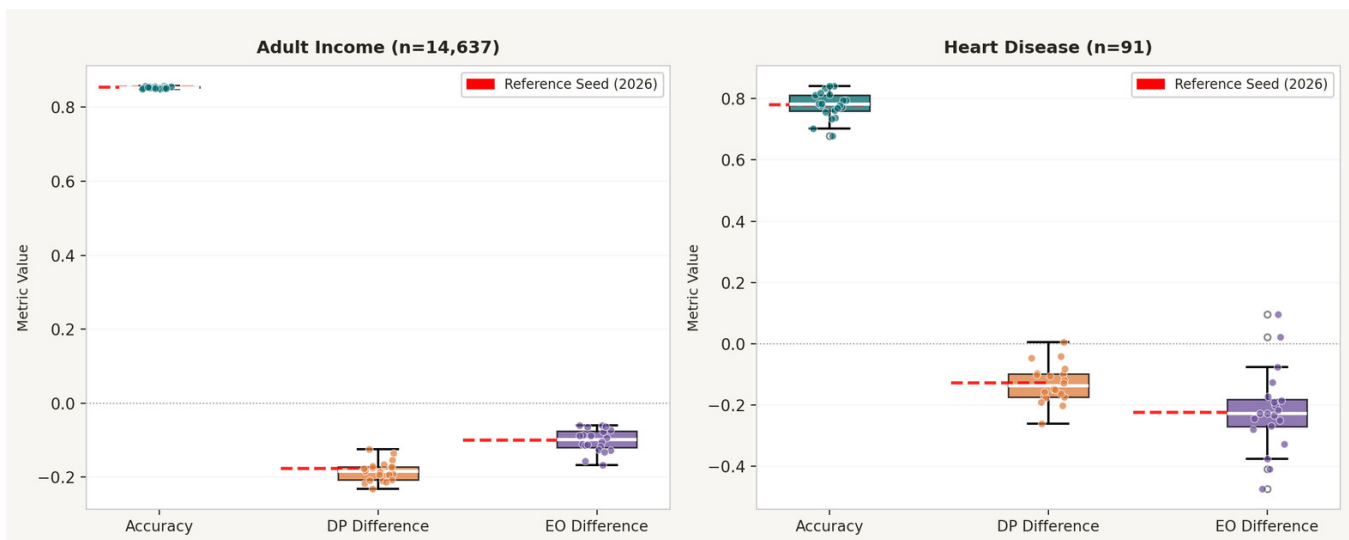


Figure 3: Distribution of accuracy, DP difference, and EO difference across random seeds for Adult Income and Heart Disease. The Heart dataset shows substantially greater variability, especially for fairness metrics.

In practice, fairness-aware optimization should not be deployed as an automated decision mechanism without human oversight and domain governance. In healthcare settings, fairness metrics should complement rather than replace clinical judgment, while in socioeconomic applications, fairness interventions should be evaluated alongside broader institutional and policy considerations.

In the Adult Income dataset, the main societal concern is unequal access to favorable economic predictions. If a predictive model assigns positive outcomes more frequently to one group than another, it may reinforce existing inequalities in employment, lending, or income-related decision-making. The improvement in demographic parity observed

in this study suggests that fairness regularization can help reduce group-level disparities without necessarily reducing accuracy. However, this does not mean that demographic parity is always the correct ethical target. In some settings, equalizing outcome rates may hide other forms of unfairness, especially when groups differ in base rates, data quality, or historical access to opportunity.

In the Heart Disease dataset, the ethical stakes are different because prediction errors can correspond to missed diagnoses. Equal opportunity is especially relevant in this setting because it focuses on whether individuals who truly have the condition are identified at similar rates across groups. A large equal opportunity gap means that one group may

experience more false negatives, which can delay treatment and worsen health outcomes. Prior research on algorithmic fairness in healthcare shows that unfair AI systems can contribute to unequal diagnosis, treatment, and billing outcomes, particularly for underrepresented groups (Chen et al. 2023). Therefore, even small differences in fairness metrics may have serious consequences when the model is used in a clinical or health-related context.

The cross-domain differences found in this study also show why fairness regularization should not be applied as a fixed rule. The same regularization strategy produced strong improvements in the Adult Income dataset but more limited improvements in the Heart Disease dataset. This suggests that fairness interventions are shaped by dataset size, feature structure, subgroup composition, and the relationship between the sensitive attribute and the target variable. From an ethical perspective, this means that model developers should not report a single accuracy score or a single fairness score as sufficient evidence of responsible deployment. Instead, fairness should be evaluated across multiple metrics, multiple random seeds, and domain-specific harm scenarios.

There are also important long-term considerations. Fairness constraints can change model behavior at deployment time, but they do not automatically guarantee better social outcomes over time. Prior work on delayed impact shows that fairness constraints such as demographic parity and equal opportunity can sometimes produce unintended long-term welfare effects depending on the decision context and feedback process (Liu et al. 2018). This is relevant to the present study because λ controls the strength of the fairness constraint. A very high value of λ may reduce one measured disparity while introducing other harms, such as reduced predictive performance, miscalibration, or poorer outcomes for the group the intervention was intended to help.

These results support the view that fairness regularization should be treated as a tunable design parameter rather than a fixed ethical requirement. Responsible use of fairness-aware machine learning requires transparency about the chosen fairness definition, justification for the selected λ value, and reporting of trade-offs between accuracy, AUC, demographic parity difference, and equal opportunity difference. This aligns with broader AI risk management guidance, which emphasizes that trustworthy AI requires balancing fairness, accountability, transparency, reliability, safety, and other system-level risks rather than optimizing a single metric in isolation (National Institute of Standards and Technology 2023).

Overall, the ethical implication of this study is that fairness-aware optimization can support more responsible model development, but it cannot replace domain judgment. In income prediction, fairness analysis should consider how model outputs may affect access to economic opportunity. In healthcare prediction, fairness analysis should prioritize the consequences of false negatives and unequal diagnostic sensitivity. Across both domains, fairness metrics should be interpreted as tools for ethical evaluation rather than as complete definitions of justice.

Limitations

This study is restricted to logistic regression, a linear model that may not capture non-linear fairness dynamics present in gradient boosting trees or neural networks. The study evaluates only tabular datasets and linear models, so conclusions may not directly generalize to foundation models or large-scale neural systems. Because each domain in this study is represented by a single dataset, the observed cross-domain differences may partly reflect dataset-specific properties, such as sample size and feature richness, rather than domain-level effects in general; disentangling these factors would require evaluating multiple datasets within each domain. Relatedly, because λ is tuned separately for each dataset, this study cannot directly confirm whether recommended λ values transfer within a contextual domain. Theoretically, this is unlikely: Liang et al. (2024) show that a dataset property termed group-balance, rather than domain category itself, determines the shape of the fairness–accuracy frontier, suggesting that λ is better understood as a function of dataset-level statistical properties (e.g., group-size ratio, label prevalence, sample size) than of domain membership. Confirming this would require applying the same λ sweep to additional datasets within each domain, which we leave to future work. The sensitive attribute is simplified to a binary male/female variable, which cannot fully represent the complexity of demographic identity in practice. The fairness penalties are probability-based soft proxies during training, while final metrics are computed on thresholded binary predictions, so the optimized objective does not perfectly align with reported metrics. Only two fairness definitions are examined, and alternative criteria such as equalized odds, predictive parity, and individual fairness may yield different conclusions. The study also does not address long-term or dynamic fairness impacts, where static constraints applied at training time may not prevent cumulative disparities across repeated deployments.

Conclusion

This study investigated fairness–accuracy trade-offs in logistic regression models trained with gradient descent under fairness regularization. By introducing a tunable parameter λ , the study analyzed how fairness constraints affect predictive performance and group fairness across two domains and two fairness definitions.

The results demonstrate that incorporating fairness directly into the optimization objective is effective and stable across the tested settings. For Adult Income, demographic parity regularization at $\lambda \approx 0.8$ reduces the DP gap by over 43% with no accuracy loss, showing that fairness and predictive performance are not necessarily in conflict when training data is large and feature-rich. In contrast, Heart Disease reveals the limits of regularization under constrained data conditions: EO regularization produces negligible improvement, while DP regularization requires careful tuning around $\lambda \approx 0.5$ to avoid unnecessary accuracy degradation. Across both domains, fairness metrics are substantially more sensitive to random seed variation than accuracy metrics, establishing multi-seed evaluation as a necessary practice for

reliable fairness research.

Future work should extend this analysis to non-linear and black-box architectures, including gradient-boosted trees and deep neural networks, explore multi-attribute sensitive variables, investigate adaptive regularization that adjusts λ dynamically during training, and examine long-term fairness impacts under the constraint frameworks studied here. Overall, the study highlights that fairness-aware optimization is not a universally transferable solution, but a context-dependent design process requiring empirical validation, ethical justification, and domain-sensitive evaluation.

References

- Blum, A.; and Stangl, K. 2019. Recovering from Biased Data: Can Fairness Constraints Improve Accuracy? ArXiv:1912.01094 (last revised 2024).
- Brzezinski, D.; Stachowiak, J.; Stefanowski, J.; Szczech, I.; Susmaga, R.; Aksenyuk, S.; Ivashka, U.; and Yasinskyi, O. 2025. Properties of Fairness Measures in the Context of Varying Class Imbalance and Protected Group Ratios. *ACM Transactions on Knowledge Discovery from Data*.
- Buyl, M.; Defrance, M.; and De Bie, T. 2024. fairret: A Framework for Differentiable Fairness Regularization Terms.
- Chen, R. J.; Wang, J. J.; Williamson, D. F. K.; Chen, T. Y.; Lipkova, J.; Lu, M. Y.; Sahai, S.; and Mahmood, F. 2023. Algorithm Fairness in Artificial Intelligence for Medicine and Healthcare. *Nature Biomedical Engineering*.
- Defrance, M.; and De Bie, T. 2025. Maximal Combinations of Fairness Definitions. *Journal of Artificial Intelligence Research*.
- Dwork, C.; Hardt, M.; Pitassi, T.; Reingold, O.; and Zemel, R. 2012. Fairness Through Awareness. In *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*, 214–226.
- Hardt, M.; Price, E.; and Srebro, N. 2016. Equality of Opportunity in Supervised Learning. In *Advances in Neural Information Processing Systems*, volume 29.
- Hort, M.; Zhang, J. M.; Sarro, F.; and Harman, M. 2024. Search-Based Automatic Repair for Fairness and Accuracy in Decision-Making Software. *Empirical Software Engineering*, 29(1).
- Kpatcha, E. 2025. Balancing Fairness and Accuracy in Machine Learning-Based Probability of Default Modeling via Threshold Optimization. *Journal of Risk and Financial Management*, 18(12): 724.
- Liang, A.; Lu, J.; Mu, X.; and Okumura, K. 2024. Algorithm Design: A Fairness-Accuracy Frontier.
- Liu, L. T.; Dean, S.; Rolf, E.; Simchowitz, M.; and Hardt, M. 2018. Delayed Impact of Fair Machine Learning. In *Proceedings of the 35th International Conference on Machine Learning*.
- Maheshwari, G.; and Perrot, M. 2023. FairGrad: Fairness Aware Gradient Descent.
- Maritn, D.; and Ogbuokiri, B. 2026. Reducing Representation Bias through Fairness-Driven Sampling in Contrastive Learning. In *The 39th Canadian Conference on Artificial Intelligence*, 1137–1144. PMLR.
- Mehrabi, N.; Morstatter, F.; Saxena, N.; Lerman, K.; and Galstyan, A. 2022. A Survey on Bias and Fairness in Machine Learning. *ACM Computing Surveys*, 54(6): 1–35.
- National Institute of Standards and Technology. 2023. Artificial Intelligence Risk Management Framework.
- Rahman, H.; and Ogbuokiri, B. 2026. Uncovering Latent Subgroups: Spectral Clustering for Fairness Analysis in Contrastive Embeddings. In *The 39th Canadian Conference on Artificial Intelligence*, 1076–1083. PMLR.
- Rezaei, A.; Fathony, R.; Memarrast, O.; and Ziebart, B. 2020. Fairness for Robust Log Loss Classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34.4, 5511–5518.
- Rychener, Y.; Taskesen, B.; and Kuhn, D. 2024. Metrizing Fairness.
- Sánchez-Marqués, R.; Garcia, V.; and Sanchez, J. S. 2026. Addressing the Balance Between Fairness and Performance in Glioma Grade Prediction Using Bias Mitigation Techniques. *Scientific Reports*.
- Sufian, M. A.; Alsadder, L.; Hamzi, W.; Zaman, S.; Sagar, A. S. M. S.; and Hamzi, B. 2024. Mitigating Algorithmic Bias in AI-Driven Cardiovascular Imaging for Fairer Diagnostics. *Diagnostics*, 14(23): 2675.
- Taskesen, B.; Nguyen, V. A.; Kuhn, D.; and Blanchet, J. 2020. A Distributionally Robust Approach to Fair Classification.
- Trotter, C.; and Chen, Y. 2024. Exploring Fairness-Accuracy Trade-Offs in Binary Classification: A Comparative Analysis Using Modified Loss Functions. In *Proceedings of the ACM Conference on Fairness, Accountability, and Transparency*.